上下文过滤缓解政治意识形态预测中的稀缺性与选择偏差

Chen Chen*[1]、Dylan Walker†[2] 和Venkatesh Saligrama‡[3]

[1]管理与经济学院，香港中文大学（深圳）[2]阿吉罗斯商学院与经济学院，查普曼大学[3]工程学院，波士顿大学

2023年2月2日

arXiv:2302.00239v1 [cs.LG] 1 Feb 2023

摘要

我们提出了一种新颖的政治意识形态预测（PIP）监督学习方法，该方法能够预测分布外输入。提出这一问题的动机在于，人工数据标注成本高昂，而自我报告标签往往稀缺，并且存在显著的选择偏差。我们提出了一种新颖的统计模型，将文档嵌入分解为两个向量的线性叠加：一个与意识形态无关的潜在中立语境向量，以及一个与意识形态一致的潜在立场向量。我们训练了一个端到端模型，其输出包括中间层的上下文向量和立场向量。在部署时，我们的模型仅利用预测得到的立场向量，为输入文档预测标签。在两个基准数据集上，我们证明，即使仅使用少至5%的有偏数据进行训练，我们的模型仍能输出预测结果，且准确率显著高于当前最先进的方法。通过众包，我们验证了上下文向量的中立性，并表明上下文过滤会产生意识形态集中现象，从而能够对分布外示例进行预测。

1 引言

政治意识形态预测（PIP）受到多种应用需求的推动，例如了解政府的政策制定（Potrafke，2018年；Hunt，2009）、立法者的党派或非党派行为（Zingher，2014；Nice，1985年；Berry et al.，2007）、普通公众对立法的情感支持（Berry et al.，2007；Budge et al.，1987；Rudolph，2009）等。

*chenchen2020@cuhk.edu.cn
†Dylan@chapman.edu
‡srv@bu.edu

我们的总体目标是从书面文本、媒体帖子和演讲中推断立法者或普通公众的意识形态。我们可以采用标准的监督学习方法，使用真实标签（例如自由派与保守派）训练 PIP 模型。在某些情况下，例如针对立法者，可以根据政党归属信息推断真实标签，而这些信息一直被用作意识形态的良好替代指标（Poole and Rosenthal，2001）。

选择偏差。 在其他情形下，例如社交媒体帖子的领域，真实值难以获取，且依赖于用户自我报告的政治立场（Bakshy 等人，2015）。这类自我报告的政治立场通常很少见，而且一旦报告，往往呈现极端倾向（Conover 等人，2011）。虽然可以利用标注人员（例如亚马逊土耳其机器人）进行人工标注（Preo,tiuc-Pietro 等人，2017），但由于过度抽样了极端且活跃的发帖者所发布的帖子，选择偏差仍然存在（Cohen 和 Ruths，2013；Mustafaraj 等人，2011；Kohut 和 Bowman，2018；Gao 等人，2015）。

理由。 仅使用稀缺且极化群体的数据训练的黑箱模型，很可能无法很好地预测那些观测不足的温和派；他们是多数群体，并且不愿在网上透露或讨论自己的政治立场（McClain，2021；Kohut 和 Bowman，2018；Gao 等人，2015）。推断多数群体的观点，有助于我们理解他们对立法／行政行动的支持程度（Craig 和 Richeson，2014），或文化群体的实际影响力（Bisin 和 Verdier，2000）；近期堪萨斯州的投票便是一例，政策制定者忽视了沉默／不太积极发声的多数群体的声音，因而高估了公众对堕胎禁令的支持（Lysen 等人，2022）。

所提方法。 为应对稀缺性和选择偏差，我们提出了一种新方法，该方法

在训练过程中强制将文本嵌入分解为语境和位置，并在具有双模态先验分布的改进型变分自编码器框架中训练深度神经网络模型。

语境过滤。 我们提出将文档嵌入分解为两个潜在向量的线性（正交）叠加，这两个向量分别由上下文向量和位置向量构成。我们训练一个采用这种分解方式的深度神经网络模型，以确保上下文向量不包含意识形态相关信息，而通过过滤语境得到的残差位置向量则独自承载所有意识形态信息。为避免出现平凡解，我们要求所有文档的上下文向量都属于一个低维流形。我们的观点是，存在一组稀疏的常见词（例如枪支、移民、税收等），其不带立场的成分用于为文档提供语境，而这些词对应的嵌入构成一个低维空间。

去噪。 在数据稀缺的情况下，文档呈现出显著差异，使得意识形态难以辨别。这是因为稀缺的文档集合分布于多种不同主题之中。不含语境的位置向量会抑制导致文档嵌入之间出现巨大差异的噪声，而这些差异与 PIP 的相关性很低，从而使意识形态表征更容易学习。

主要实验发现

最先进水平预测。我们的模型 BBBG 在标签稀缺和存在偏差的各种情形下，其意识形态预测均全面优于先前研究。此外，与 BERT/RoBERTa 等其他嵌入相比，GloVe 的预测在标签稀缺/有偏的情形下尤其更为准确。

潜在上下文的中立性。众包实验表明，上下文向量在不同主题中与中性词/短语相关联。

意识形态纯度。上下文过滤会导致意识形态净化，并提升预测效果。

知识迁移。BBBG 能够有效地将知识迁移到域外场景，例如：(a) 将意识形态预测从极端两极化作者推广到接近温和派的作者；(b) 将文档从“已见主题”推广到“未见主题”文档。

2 相关工作

我们介绍了专注于利用机器学习从文本中预测意识形态的先前研究。

监督学习。（Gentzkow 和 Shapiro，2010年）提出了一种基于学习的方法，用于在国会报告语料库上进行意识形态预测。针对同一数据集的其他研究，在相同设定下采用了现代学习方法（Pla 和 Hurtado，2014；Gerrish 和 Blei，2011；Iyyer 等人，2014）。利用社交媒体数据（推文、论坛帖子）的研究旨在将公众用户映射到自由派—保守派光谱上，或直接预测其政党标签（Levendusky 和 Pope，2010年；Baldassarri 和 Gelman，2008）。

文本特征。 上述研究大多采用现成的预训练文本表示，例如词袋、词频—逆文档频率、语言查询与词计数，或词嵌入——GloVe、BERT 或 RoBERTa（Preo. tiuc-Pietro 等人，2017；Conover 等人，2011；Mokhberian 等人，2020；Liu 等人，2019年）。

稀缺且有偏的数据。作为监督信号的标签来自用户自我报告的党派归属（例如民主党人或共和党人），或由标注人员进行人工标注（Conover 等人，2011；Preo.tiuc-Pietro 等人，2017；Cohen 和 Ruths，2013），因此本质上受到标签稀缺和选择偏差的影响。例如，大多数公众不愿透露自己的党派归属，也不愿在线参与政治话语（Bakshy 等人，2015；McClain，2021）。此外，为研究观点或意识形态而提出的带标签文本收集方法，往往会过度抽样“发声的少数派”，同时忽视或减少对“不太积极发声的多数群体”的抽样（Moe 等人，2011；Kohut 和 Bowman，2018；Mustafaraj 等人，2011；Gao 等人，2015）。人工标注受到人员配置以及标注人员自身认知偏差的严重制约（Yu 等人，2008；Xiao 等人，2020）。最后，明显表达观点的用户与温和派用户生成的内容之间存在显著的领域差异（Cohen 和 Ruths，2013）。出于多种原因，用于学习意识形态的模型通常泛化能力较差（Cohen 和 Ruths，2013；Yan 等人，2017，2019）。

通用方法。 迄今为止，鲜有研究在基于文本数据训练意识形态预测模型时，同时考虑标签稀缺和选择偏差。对于 Twitter 数据，研究者会在图学习框架中利用用户的社交互动（例如，提及/关注关系）来解决标签稀缺问题。然而，这些方法

可能不适用于缺少社交互动和连接，或无法观测到这些信息的场景（Xiao 等人，2020年）。传统上，半监督学习（SSL）可以应对标签不足的问题（参见（Ortigosa-Hernández 等人，2012a年； Oliver 等人，2018； Kingma 等人，2014））， 而领域自适应（DA）（Daumé III，2009； Glorot 等人，2011）可以处理训练样本与测试样本之间的分布差异，从而有望为解决选择偏差提供一种方案。我们在实验中比较了这些替代方法、词嵌入以及其他先前研究。

3 文本生成的统计模型

文档分解。 我们首先将文档编码到合适的嵌入空间（GloVE、RoBERTa 等）中，并令 $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{\mathcal{D}}$ 表示该嵌入，作为学习器的输入。我们将$\mathbf{x}$分解为两个主要部分：

$$\mathbf{x} = \mathbf{c}(\theta) + \mathbf{f}(\theta, z) + \epsilon \quad (1)$$

其中，$\mathbf{z} \in \mathbb{R}^M$ 是$\mathbf{x}$的潜在意识形态向量，向量$\mathbf{c}$ 和$\mathbf{f}$ 分别表示中性语境和过滤后的位置向量分量。ϵ 是一个随机向量，用于捕捉文档间的个体特异性变化。参数 θ表示作者对主题的选择，例如“枪支”或“堕胎”。然而，若无进一步约束，这种分解是不可识别的。我们对上下文向量施加低维结构，并在训练过程中将上下文向量与意识形态内容区分开来。

语境向量的低维中性编码。 我们基于以下直觉展开。对于某个党派主题θ_i，假设存在一条主要的极化轴 pa_{θ_i}，其具体方向取决于该主题；不同意识形态群体（例如自由派或保守派、左派或右派）在发表演讲时[1]会沿此轴呈现差异。原则上，我们提出让语境与极化轴 pa正交。实践中，由于 pa 是先验不可观测的，我们通过谨慎初始化来施加这一约束，并在训练期间根据经验将位置$\mathbf{f}$ 与极性对齐，同时验证事后正交性。

初始化。 为了初始化上下文向量，我们采用以下流程。首先，我们通过生成一组“中性”种子词来确定主题和短语（以下统称为种子），这些种子可以通过人工构造，或选择频率高且具有较低 χ^2值的词语获得（参见附录第 A.4节）；随后，通过扩展至种子的邻域来补充这组词语，从而为每个主题生成一组种子 i， a_{ij}。主题初始化为种子词的汇总嵌入 $T_{0i} = \sum_j a_{ij}$。上下文向量初始化为主题的混合， $c_0(\theta_0) = T_0\theta_0$。最终，主题的选择， θ 以及主题本身T 都会在训练过程中学习，初始状态来自初始化 $T = T_0$ （更多细节和其他初始化方法见附录 A.4 和 A.5）。低维性则源于主题矩阵为低秩矩阵。这种方法有效去除了文档嵌入之间差异较大的方差，而这类方差对意识形态学习和预测的相关性较低，因此使意识形态表征的学习更加容易。

多模态先验。 由于个人明确表明自己在党派归属上认同具有共同意识形态立场的群体是很常见的现象，我们假设意识形态表征 $\mathbf{z}$ 来自多模态先验分布。在美国政治领域，一个合理的模态数是二，这表示两党制／极化。在国会和政治辩论论坛的语境下，这一假设得到了以往研究的支持（PooleandRosenthal，2001； Bonica，2014），并反映了美国社会中的普遍趋势 [2]。不过，我们的框架可以推广到多个模态。

4 方法

数学背景 除了第 3 节中的记号外，我们用 $\mathbf{y} \in \mathcal{Y}$ 表示输出意识形态标签，其取值属于一个有限集合。令$p_s(\mathbf{x}, \mathbf{y})$ $p_t(\mathbf{x}, \mathbf{y})$分别表示源域和目标域的联合分布（在语境明确时，我们省略这些下标）。学习器可以访问一个有标签集 $\mathcal{D}_s$ ，其中包含 N_s 个独立同分布数据点 $(\mathbf{x}_i, \mathbf{y}_i) \sim p_s$ ，以及一个无标签集 $\mathcal{D}_t$，其中包含 N_t 个独立同分布的目标域输入实例 $\mathbf{x}_i \sim p_t$。

模型输入与输出。 学习器的目标是根据给定的目标域输入，预测目标域的标签$\hat{\mathbf{y}}$ 。此外，我们训练的模型还会为每个输入文档输出语境向量$\mathbf{c}$、主题选择 θ、位置向量$\mathbf{f}$和意识形态向量$\mathbf{z}$。

训练损失。 我们的损失函数由以下部分组成：重 con-

[1]从数学角度来看，这可以通过第一主轴进行近似；或者用一种更简单的方法，计算自由派群体的平均嵌入与保守派群体的平均嵌入之差。

[2]https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/

构损失，作用于输入$\mathbf{x}_i \in \mathcal{D}_s \cup \mathcal{D}_t$；以及交叉熵损失 $CE(\mathbf{y}_i, \hat{\mathbf{y}}(\mathbf{z}_i; \gamma))$，作用于 $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}_s$；$\hat{\mathbf{y}}$ 是一个由参数 γ 控制的网络的 Softmax 输出，其输入为编码后的潜在表示$\mathbf{z}_i$（见下文）。

重构损失。我们的出发点是优化边缘分布，即 $\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})} \log p(\mathbf{x})$，但由于这难以处理，我们通过放松条件来得到证据下界（Kingma 和 Welling，2013）：

$$L_R(\eta, \phi, \lambda) \triangleq \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\eta(\mathbf{x}|\mathbf{z}) + \ln p_\lambda(\mathbf{z}) - \ln q_\phi(\mathbf{z}|\mathbf{x})]] \quad (2)$$

其中，$q(\mathbf{x})$ 是源域和目标域输入的经验分布；$q_\phi(\mathbf{z}|\mathbf{x})$ 是编码器，$p_\eta(\mathbf{x}|\mathbf{z})$ 是解码器；$p_\lambda(\mathbf{z})$ 是先验分布。ϕ、η、λ 分别是它们的参数。现在，我们调用公式 1 中的统计模型，将 x 分解为 f 和 c。根据上下文向量的中立性，$p(\mathbf{c}|\theta, \mathbf{z}) = p(\mathbf{c}|\theta)$；又由于在给定$\theta$的条件下，c 和 f 相互独立，因此可得：

$$L_R(\eta, \phi, \lambda, \theta) \triangleq \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p(\mathbf{f}|\mathbf{z}, \theta)] + \ln p(\mathbf{c}|\theta) + \ln p_\lambda(\mathbf{z}) - \ln q_\phi(\mathbf{z}|\mathbf{x}) + \ln p(\epsilon|\mathbf{c}, \mathbf{f})] \quad (3)$$

公式 3 中的第一项对应于生成 f 的解码器；第二项描述从 θ 生成 c。第三项 $p_\lambda(\mathbf{z})$ 是 z 的先验分布。最后一项对应于通过均方损失最小化重构误差（$\epsilon \triangleq \mathbf{x} - (\mathbf{f} + \mathbf{c})$）。与传统的变分自编码器不同，我们采用多模态先验，并按照（Tomczak 和 Welling，2018）中的方法将其建模为 K 模态高斯分布。我们通过采样并将K 伪输入馈入同一个编码器来近似先验分布，即 $p_\lambda(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K} q_\phi(\mathbf{z}|\mu_n)$。这里，$K$ 是指定先验模态数的超参数。对于美国政治领域，我们选择 $K = 2$，但我们的方法可以推广到具有不同 K的设置。

深度神经网络（DNN）训练。我们通过优化总损失函数 $L = L_R(\eta, \phi, \lambda) + \mathbb{E}_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_S} CE(\mathbf{y}, \hat{\mathbf{y}}(\mathbf{z}; \gamma))$，对所有参数（$\eta$，$\phi$，$\lambda$，$\gamma$）进行反向传播，从而端到端地训练 DNN；同时在 z 上使用双模态先验分布，并将最终得到的预测器称为双分支双模态高斯变分自编码器（BBBG）。我们的模型与传统 VAE 的不同之处在于，它使用多模态先验分布，并在模型生成部分中通过两个分别学习的组件重构输入（见附录图 5、6）。

单分支消融。作为消融实验，为了理解式1的重要性，我们通过删除模型中用于学习中性语境的分支，放宽中性约束，并将这一算法称为 SBBG （单分支变体）。在简化版本中，不再估计语境，因此模型退化为采用高斯混合先验分布的标准 VAE 框架，但仍使用相同的 K模态 VampPriors。实现和训练细节见附录 B.2。**监督的合理性。**传统 VAE 模型采用无监督学习。但由于过度正则化，它们往往表现不佳（Dilokthanakul 等人，2016）。在我们的两项实验中，无监督 VAE 变体似乎都会坍缩为单模态高斯先验。为解决这一问题，我们通过少量带标签样本提供监督，使模型收敛到双模态先验分布。这意味着在训练期间，我们允许模型看到一些意识形态标签，并将意识形态预测损失反向传播，以帮助调整编码器中的参数。事实证明，有效训练此类系统所需的标签数量为总样本的 5∼8%。

5 实验

我们使用两个基准数据集进行实验： 国会议员演讲 和 枪支辩论论坛，以此将所提 BBBG 方法与知名的先验机器学习方法进行比较。我们在第 A.1 节详细介绍这些数据集，并指出选择它们的原因是需要确定文本和作者的真实极端程度（从极端到中立）。其他数据集都是二元的，因此不提供此类信息。我们所选的数据集允许对具有不同意识形态极端程度的子群体进行子抽样，从而能够在选择偏差下验证所提方法。

模拟标签稀缺与选择偏差。我们在数据集中屏蔽真实标签，以模拟标签稀缺和选择偏差。我们将监督水平定义为未屏蔽样本的百分比，Sup 该比例决定了在稀缺性或稀缺性与偏差（偏向极端程度）并存的情况下屏蔽的程度。为模拟稀缺性，我们随机屏蔽了 $(1 - Sup)$% 的数据。为模拟稀缺性与选择偏差，我们屏蔽了极端程度最低的作者所对应的 $(1 - Sup)$% 数据。我们将前一种做法称为无偏监督 ，

后一种做法称为有偏监督。屏蔽样本的预测损失不会用于随机梯度下降（SGD）以更新网络权重。关于屏蔽流程和超参数调优的详情，请参见第 A.2、A.6节。

按主题分类。我们旨在揭示 BBBG 的其他输出所发挥的作用，即语境向量 $\mathbf{c}$、过滤后的位置向量 $\mathbf{f}$以及意识形态向量 $\mathbf{z}$。为此，出于消融分析的目的，我们手动将国会数据集划分为 68 个主题（请注意，BBBG 训练并不依赖于我们的分类方式），其中约有 25 个党派主题和 43 个非党派主题（见第 A.7 节）。我们开展了多项实验，研究上下文中立性、通过上下文过滤进行去噪，以及向未见主题进行知识迁移。

基线方法。以往方法利用预训练的文本特征嵌入，并以各种方式使用有监督和无监督数据。GloVe 嵌入是我们的主模型（BBBG）和其他基线模型的默认嵌入。基于 BERT 或 RoBERTa 的模型会在名称中注明。所有模型的完整实现细节和比较见第 A.2 节。

仅使用带标签数据训练的方法。其中包括 GS（Gentzkow 和 Shapiro，2010年），以及标准机器学习方法，例如支持向量机、随机森林、XGBoost 和 8 层深度神经网络（8l-DNN），其容量与 BBBG 和门控循环单元相近。这些方法依赖预训练文本特征，例如 GloVe 或 RoBERTa（Dey 和 Salem，2017；Liu 等人，2019年）。同时利用带标签和无标签目标域数据的方法。其中包括半监督学习方法，例如基于 k 近邻的标签传播（LS-KNN）（Ortigosa-Hernández 等人，2012b年），以及与深度学习相结合的自训练方法（ST-DNN），其容量与 BBBG 相近；还包括集成学习方法，例如随机森林（ST-RF）（Yarowsky，1995；Tanha 等人，2017；Zhang 等人，2014）。ST 基于迭代式伪标签（Zou 等人，2018）；最后，还采用基于 RoBERTa 嵌入的领域自适应方法（RoBERTa-DA）来处理领域偏移（Glorot 等人，2011）。

5.1 国会数据集上的预测

在稀缺且有偏的情形下，BBBG 的表现优于先前研究。在基线模型中表现最好的是包括深度学习方法在内的模型、半监督方法（LS-KNN 或 ST-DNN）在结果标签充足且监督抽样无偏时表现良好（其他模型，如 BERT-DNN、(Devlin 等人，2018)则表现欠佳（见 B.1 节））。然而，随着稀缺性增加，其表现显著下降。如表 1和表 6所示，一旦监督比例低于 20%，BBBG 的表现便显著优于所有基线模型；无论监督有偏还是无偏，随着监督比例降低，这一差距都会扩大。即使在监督抽样有偏时仅使用 5% 的标签（~10k），或在监督无偏时仅使用 3%（~6.6k），BBBG 也能高精度地预测作者的政党标签。

GloVe 与 BERT 和 RoBERTa 的比较[3] 首先要注意，当稀缺性／偏差不是问题时，采用 GloVe 的 BBBG 与采用其他最先进的语言模型 BERT 和 RoBERTa 的 BBBG 表现相近（表 1，第 B.1 节）。然而，在标签稀缺（ <8%，约为 1.76 万）且存在选择偏差的情况下，无论与深度神经网络还是 SBBG 框架结合，RoBERTa 嵌入的表现都不比 GloVe 更好或更差。这表明，RoBERTa 或 BERT 等更复杂的模型可能更依赖充足且高质量的标签，因此更容易受到监督质量不佳的影响。此外，领域适配（基于 RoBERTa 嵌入）似乎并不比其他一些基线方法更有优势，而且在几乎所有条件下都显著逊于 BBBG。总的来说，这些结果表明，我们的 BBBG 模型在应对标签稀缺和选择偏差方面都具有显著更强的韧性。

5.2 GunDebate 论坛数据集

在数据稀缺、有偏的情况下，以及在所有监督机制下，BBBG 的表现均优于先前研究。在数据稀缺（<8%，约 4.8k 份文档）和极端偏差的情况下，我们比较了 BBBG 与几种替代模型的表现。除了上述模型，我们还将其与 BERT-Sequential 进行了比较（Devlin 等人，2018；Baly 等人，2020年）。

与先前方法相比，BBBG 在枪支辩论论坛数据集上的表现显著更优。这可能是因为与国会用户相比，论坛用户的说话方式异质性更高

[3] GloVe 的一种自然初始化形式是“单词”，这有助于估计语境并进行语境过滤。我们以 SBBG 为基线模型来评估 BERT 和 RoBERTa，因为注意力机制会使它们的嵌入与周围内容交织在一起，因此难以采用单词初始化。

表1：在无偏/有偏监督下，对国会演讲数据进行政党预测的准确率，展示了排名前三的基线方法与主模型 BBBG 以及单分支变体 SBBG 之间的比较结果。最佳结果以**粗体** 显示，次佳结果带下划线。蓝色标记表示在标签稀缺时，BBBG 的表现显著优于大多数其他模型。所示百分比为三次独立试验的平均值。

	无偏监督								有偏监督							
	80%	60%	40%	20%	8%	5%	3%	1%	80%	60%	40%	20%	8%	5%	3%	1%
8l-DNN	93.2%	**94.4%**	92.6%	88.1%	84.9%	55.0%	53.6%	50.4%	84.5%	**84.8%**	86.9%	**87.9%**	65.8%	61.1%	56.7%	50.7%
GS	79.1%	79.9%	81.4%	82.3%	79.0%	77.1%	78.7%	70.4%	70.4%	71.5%	74.5%	72.0%	69.5%	71.1%	68.4%	60.7%
LS-KNN	90.2%	92.6%	92.3%	91.5%	88.8%	85.8%	82.5%	79.7%	84.7%	83.5%	83.0%	81.4%	78.2%	75.7%	71.3%	67.8%
ST-DNN	**93.5%**	94.3%	94.0%	93.0%	88.0%	77.3%	75.3%	61.6%	81.3%	81.6%	83.2%	80.8%	71.5%	63.8%	56.3%	51.9%
RoBERTa-GRU	92.8%	93.3%	83.3%	78.2%	69.9%	51.6%	51.7%	51.9%	83.6%	84.7%	84.0%	78.1%	53.1%	49.5%	50.7%	50.9%
RoBERTa-DA	90.6%	89.1%	89.4%	79.3%	75.1%	62.9%	61.8%	51.4%	84.0%	81.0%	81.9%	68.5%	69.3%	54.8%	51.4%	49.8%
RoBERTa-SBBG	93.3%	91.1%	94.0%	89.2%	76.5%	61.1%	52.7%	51.3%	**86.1%**	84.7%	**87.5%**	85.9%	51.3%	51.2%	48.6%	48.8%
SBBG	92.9%	90.4%	90.6%	89.9%	86.6%	86.7%	84.0%	74.1%	78.3%	81.3%	83.6%	82.0%	79.8%	77.5%	73.5%	65.7%
BBBG	92.7%	92.9%	**94.1%**	**93.2%**	**91.6%**	**89.8%**	**87.2%**	**81.2%**	81.3%	81.3%	85.2%	83.3%	**83.6%**	**84.0%**	**76.4%**	**68.6%**

表 2：在无偏/有偏监督下预测意识形态标签的准确率。枪支辩论论坛 数据集展示了三种基线方法、主模型 BBBG 与单分支变体 SBBG 的对比结果。最佳结果以**粗体** 显示，次优结果带下划线。BBBG 在稀缺标签条件下显著优于大多数其他模型，以蓝色标记。所示百分比为三次独立试验的平均值。

	无偏监督								有偏监督							
	80%	60%	40%	20%	8%	5%	3%	1%	80%	60%	40%	20%	8%	5%	3%	1%
8l-DNN	72.1%	72.3%	67.4%	61.0%	60.9%	61.2%	61.4%	61.3%	69.9%	69.8%	66.2%	64.8%	61.0%	61.0%	61.2%	61.2%
LS-KNN	66.2%	65.0%	64.3%	64.8%	64.9%	63.9%	62.7%	60.8%	62.4%	60.8%	61.2%	63.4%	61.7%	62.5%	61.3%	61.0%
ST-DNN	70.9%	68.6%	66.8%	64.1%	<60.0%	<60.0%	<60.0%	<60.0%	69.4%	64.8%	64.6%	65.4%	61.9%	61.2%	61.3%	<60.0%
RoBERTa-GRU	76.3%	75.7%	75.9%	71.0%	63.7%	62.2%	61.2%	59.8%	77.6%	73.8%	71.5%	68.9%	67.5%	64.2%	61.1%	58.1%
RoBERTa-DA	72.1%	73.0%	70.3%	66.4%	64.5%	61.5%	62.9%	60.5%	70.3%	73.7%	67.6%	68.9%	68.0%	63.9%	64.1%	60.7%
RoBERTa-SBBG	78.9%	76.7%	73.1%	72.0%	67.8%	62.1%	61.2%	61.1%	74.6%	74.4%	71.0%	69.4%	63.9%	60.7%	60.6%	61.0%
SBBG	96.1%	96.1%	93.6%	91.6%	83.3%	72.4%	65.1%	61.3%	92.2%	88.7%	88.4%	**86.1%**	73.0%	70.5%	66.6%	60.9%
BBBG	**98.8%**	**97.9%**	**96.0%**	**91.6%**	**85.0%**	**77.3%**	**70.7%**	**61.3%**	**94.3%**	**91.0%**	**90.2%**	85.3%	**77.3%**	**74.4%**	**71.5%**	**61.2%**

而 BBBG 更能处理这种情况。我们在表2中报告了通过汇总作者层级的预测结果，得到的发帖者二元意识形态标签（自由派与保守派）。除 SBBG 消融实验外，RoBERTa-DA 在基线方法中表现最佳。对于 SBBG，只有与 GloVe 嵌入结合时表现良好，而非与 RoBERTa 结合。当监督存在偏差时，我们的模型与最佳基线模型之间的差距扩大至 19%。

5.3 上下文中立性验证

我们开展了两项实验研究，以展示推断出的潜在上下文向量的几何特征。

众包实验。 我们拟通过众包验证，语境向量 BBBG 输出的邻域中的词语与人类对中立性的判断有多一致。我们研究：1）词语与主题的感知相关性。

2）这些词在多大程度上属于自由派或保守派（我们还提供“不知道”选项）。

为此，我们根据嵌入空间中的余弦相似度，计算了与输入文档所推断的 BBBG 上下文向量最接近的十个词（第A.8节，表5）。随后，我们选取了 6 个突出的党派主题（见表5、A.8），并为每个主题从上述 10 个最近邻词中随机抽取 5 个（下文称为邻域词）。接着，我们选取典型极端词语作为各主题的参照。这些典型极端词语是通过从已知的党派新闻媒体中人工收集五个自由派和保守派词语/短语而得（见第A.8节及补充材料）。对于每个项目，我们使用李克特量表请零工劳动者对相关性（重新缩放至 [0,1]）和意识形态倾向（重新缩放至 [-1,1]）进行评分。

首先，我们注意到，人们认为邻域词和人工挑选的词都具有相关性。与上下文向量距离较近的词，其相关性得分高于 0。843±0.007 相比之下，人工收集的保守派和自由派参照词语/短语得分为 0。703±0.009 和 0。840±0.007 分别。在意识形态倾向方面，邻域词语的得分为 0。056±0.018 而保守派和自由派参照词语/短语的得分为 0。323±0.025和 -0。341±0.023分别。邻域词语显然比参照词语更中立（更接近 0）（见图1）。根据双样本 T 检验，这一差异在 0.001 水平上显著。此外，调查中上下文向量距离排名与偏离中立点（0）的程度排名之间的斯皮尔曼秩相关系数为 0.5，验证了众包评分与我们的评分一致（见第 A.8节）。

潜在上下文与残差的正交性。对于每个党派主题（见第A.7节）， θ，令 $\mathcal{D}_\theta/\mathcal{R}_\theta$ 分别表示民主党人/共和党人生成的数据文本或演讲文档。我们将极化轴定义为 θ 的$\mathbf{pa}_\theta \triangleq \mathbb{E}_{x\sim\mathcal{D}_\theta}\mathbf{f}(x) - \mathbb{E}_{x\sim\mathcal{R}_\theta}\mathbf{f}(x)$，其中$\mathbf{f}$是 BBBG 的过滤后立场分量的输出。上下文向量与极化轴之间的夹角在不同主题间取平均后约为

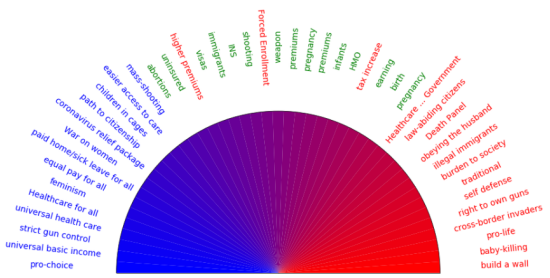


图 1：“阳光”图根据众包评分排序，展示了众包工作者如何看待党派媒体中出现的、位于 BBBG 上下文向量（绿色）邻域内的词语，并与其他刻板印象词语（蓝色和红色）进行比较。显然，这些评分与 BBBG 基本一致。少数差异似乎是众包评分噪声所致。

84 度，这表明二者近乎正交。

5.4 上下文过滤带来纯度

在此，我们测试了经过上下文过滤后，过滤后的位置向量 $\mathbf{f}$ 是否更加集中，并且与意识形态标签的对齐效果是否更好。

我们从国会（包括参议院和众议院）中选取了最活跃的 30 位发言者（17 名民主党人、13 名共和党人）所涉及的出现频率最高的 20 个主题的演讲文档样本。对于每份文档，我们收集了原始输入文本嵌入 $\mathbf{x}$ 以及相应的过滤后的位置向量 $\mathbf{f}$。我们测试了两个关键指标：(1) 第一个主成分解释的方差百分比；(2) 超过给定阈值的非零主值总数（即近似协方差矩阵的秩）。

多位作者关于不同主题的写作。我们探索了几种不同情况，以突出 BBBG 的去噪效果。我们报告的结果对应的 p 值为 0.001：

1) 作者围绕一个主题写作：在所有作者和主题上的平均值中，沿第一主轴解释的方差百分比从 0.22 增至 0.39，而排名从 53 降至 14。

2) 作者围绕多个主题写作：在所有作者上的平均值中，解释的方差百分比从 0.29 增至 0.41，而排名从 101 降至 18。

3) 具有相同意识形态标签的多位作者围绕一个主题写作：在民主党人中，解释的方差百分比从 0.20 增至 0.37；在共和党人中，则从 0.21 增至 0.35。按主题取平均后，民主党人的排名从 111 降至 18，共和党人的排名从 147 降至 31。

4) 同一意识形态标签的多位作者围绕多个主题撰写内容：民主党人的解释方差从 0.26 增至 0.40，共和党人的解释方差从 0.22 增至 0.33，而排名

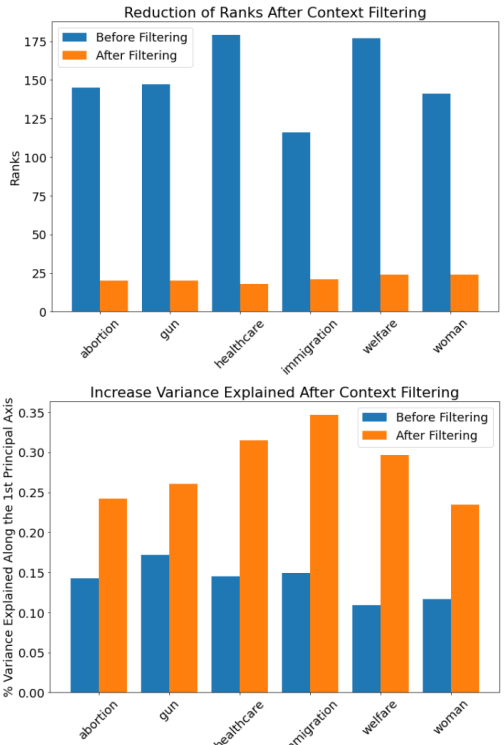


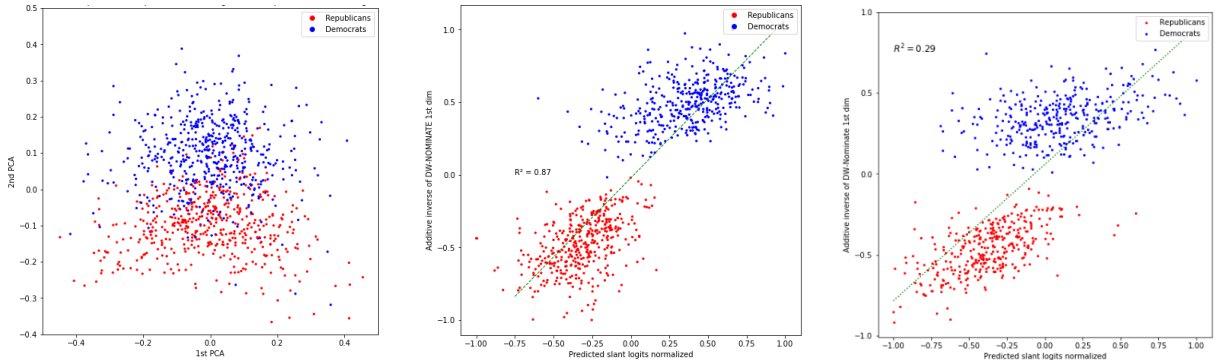
图 2：上下文过滤对意识形态预测的影响。（上）相较于未过滤编码，每个主题的位置向量集中在低维空间中。（下）相较于未过滤编码，位置向量解释的方差百分比更高。这表明，意识形态信号在少数几个维度上得到了显著放大。

民主党人的排名从 151 降至 29，共和党人的排名从 147 降至 31。

此外，我们发现，在六个党派主题中的每一个主题里，$\mathbf{f}$ 在较低秩下所解释的方差百分比显著更高（图 2）。这表明，党派主题的变化显著集中在少得多的方向上。综合来看，这些结果一致表明，上下文过滤会使信息集中到若干个“关键”方向上，并使用于预测意识形态的输入变得不那么“嘈杂”。

不同政党的过滤后的位置向量比原始嵌入分离得更明显。在保持其他方面不变的情况下，我们抽取了出现频率最高的 20 个主题。我们在先前的分析中证明，对于以残差表示的输入，第一主轴解释了至少 35% 的总变异；对于以文档嵌入表示的输入，这一比例为 20%。经过上下文过滤后，各政党意识形态中心 [4] 之间的距离增加了 5 倍（从 0.004 增至 0.02）。此外，按主题分组时，我们观察到效果存在异质性：

[4]定义为投影到第一主轴上的归一化向量的中心



(a) (b) (c)

图 3：(a) 意识形态向量 **z** 的前两个主成分表明，它捕捉了政党归属。(b&c) DW-NOMINATE 取反值（y 轴）与倾向值（x 轴）之间的相关性。(b) BBBG 的 $R^2 \sim 0.9$ 相关性更高。(c) 表现最佳的基线模型（$R^2 \sim 0.3$）

对于“堕胎”、“枪支”、“医疗保健”和“伊朗/利比亚/叙利亚”等主题，距离显著增加（超过 0.05）；而对于“文化”、“宗教”和“体育”等主题，距离的变化则微乎其微。因此，极化程度更高的主题沿第一（即主导）主轴分离得更明显。

带标签的意识形态一致性更高。为验证我们的模型是否学到了正确的意识形态表征，我们在国会演讲上进行了测试，并使用 8% 的有偏监督进行训练，观察到以下结果：

(1) 显然，意识形态不同的成员沿第二个主成分分离，这表明**z** 确实是意识形态表征（图 3a）。(2) 图 3b 展示了意识形态监督模块（图 6 ）的 logit 输出均值按作者聚合后，与 $R^2 \sim 0.9$高度相关（*DW-NOMINATE* 分数。这一数据集根据众议院和参议院成员的投票记录描述其意识形态分数，并被视为黄金标准（见第A.1节）。因此，*DW-NOMINATE* 分数不依赖国会演讲；与基线模型相比，较高的相关性意味着 BBBG 能够准确捕捉意识形态差异（图3c）。

5.5 向未见主题进行知识迁移

更好地迁移到未见主题。 在这里，我们使用监督标签训练样本，这些样本对应于属于八个主题（“信息技术”、“堕胎”、“成瘾”、“经济”、“政治行动”、“邮政服务”、“体育”、“传统能源”、“废弃物”、“劳动力”）的文档，这些文档占文档总数的 8%，并包含党派和非党派主题。随后，我们在 60 个“未见”主题上进行测试（见第 A.7 节）。如表 3 所示，BBBG 的表现优于所有其他基线方法，这表明语境过滤的位置向量能够揭示不同无关主题之间的意识形态相似性。

表 3：使用仅来自八个选定主题样本的监督标签对国会报告进行实验。此处报告了来自其他 60 个主题的文档样本的预测准确率（见第 A.7 节）。最佳结果以粗体显示，次佳结果带下划线。BBBG 的表现显著优于所有其他模型。所示百分比为三次独立试验的平均值

LS-kNN	RoBERTa -DA	RoBERTa	8l-DNN	SBBG	BBBG
63.8%	50.1%	58.2%	78.8%	<u>81.6%</u>	**86.1%**

BBBG 从极端分布迁移到近中性分布的效果更好。我们评估了 BBBG 在不同偏置分布之间泛化的有效性（见附录图4）。在这里，保守派或极端保守派发帖者约占发帖者的 45%，而自由派或非常自由派发帖者占 25%。鉴于美国大多数民众在意识形态上并不极端，但在政治上并不活跃（Wojcik 和 Hughes, 2019年），因此理解他们尤为重要（DelliCarpini，2000）。我们在枪支辩论论坛、轻度倾向发帖者这一子群体以及留出的极端发帖者集合（即训练期间屏蔽的发帖者）上评估了训练好的模型（见表 4）。报告 ST-DNN 的结果，是因为它最能代表先前研究；而如观察所示，BBBG 的表现优于 ST-DNN。其他基线模型（例如 8l-DNN）也呈现出类似结果。

6 结论

我们提出了一种新颖的深度学习方法，用于应对政治意识形态预测问题中通常会出现的标签稀缺和选择偏差。我们的方法学习将文本嵌入分解为中性语境和经过上下文过滤的位置向量，

表4：比较由不同有偏监督级别训练的模型对极端群体和轻微倾向群体意识形态标签的预测准确率

	轻微倾向		极端	
监督级别	ST-DNN	BBBG	ST-DNN	BBBG
80%	69.3%	83.1%	98.6%	99.2%
60%	63.2%	74.2%	93.4%	99.1%
40%	58.7%	68.2%	88.4%	98.5%
20%	59.4%	65.9%	81.2%	94.5%
8%	58.4%	60.1%	70.4%	85.5%
5%	50.4%	56.5%	70.6%	81.9%

其中后者包含意识形态信息。除了展示该方法在基准数据集上提高了预测准确率外，我们还通过消融实验揭示了这一方法的重要特性。我们通过众包验证了语境中立性，通过上下文过滤验证了意识形态集中度，并验证了该方法能够将知识迁移到涉及新主题的文档，以及意识形态与训练数据相似度较低的人群。展望未来，我们可以检验模型能否拓展到 Twitter 等更广泛的社交平台，或者学习并验证类似于 DW-NOMINATE 的连续意识形态表征。

7 伦理考量

7.1 数据收集与使用

国会报告语料库可公开获取，也可直接从网上下载。Debatepolitics.com 上的帖子是在严格遵循源域所有者的使用条款和隐私权的前提下收集的。参与数据收集过程的作者均已完成涉及人类受试者的研究训练。

Debatepolitics 数据将在收到请求后发布。论坛发帖者的个人信息将不会向公众公开。

7.2 BBBG 的益处与潜在误用

预期用途。本项目旨在提供一种方法，以克服意识形态预测文献中尚未得到充分解决的标签偏差和稀缺性问题。它还提供了一种有用的意识形态表示，可进一步用于其他学习目的。该方法尤其适用于预测和评估非极端群体的立场；这类群体往往不积极参与政治（参见第 4.2.3 节）。

近期堪萨斯州堕胎投票 [5] 表明了预测沉默的大多数的倾向有多么重要。缺乏这类工具，立法者更可能错误地将发声的少数派的观点推及全体人口。此外，基于发声的少数派观点得出的错误推断，可能会对政治疏离产生重大影响[6]。

然而，与任何机器学习模型一样，其真实能力存在被过度泛化的风险。评估模型的输出时，必须充分考虑输入源的特征。文本作者可能存在显著的领域差异，因此，根据对我们这组作者的研究得出的任何结论，都不能直接推广到其他群体。

滥用风险与潜在危害。除非用户以非预期的方式解读预测结果，否则我们的模型不应造成危害。该模型旨在揭示某个群体或人口的意识形态分布，而非评判个人。尽管在现实情境下，其准确度高于大多数模型，但输出并非毫无误差。对于立场略微偏向某一方的人和温和派，模型仍可能给出与真实值不符的结果。不过，与先前的最先进水平相比，我们的模型降低了这种风险。如果将模型用于针对弱势群体作出决策，其潜在危害可能会被放大。

我们的模型生成的预测和见解不应被视为事实或金科玉律。我们还建议，对任何与政治相关的研究结果都应持怀疑态度，并鼓励模型使用者在相应的应用领域进行仔细评估，查阅更多来源，或咨询政治学家以获取专业观点。

[5] https://apnews.com/article/堪萨斯州-堕胎-投票-重新计票\-e874f56806a9d63b473b24580ad7ea0c

[6] https://academic.oup.com/pa/article/68/suppl_1/241/1403570

参照词

Eytan Bakshy、Solomon Messing 和 Lada A Adamic。2015。脸书上接触意识形态多元的新闻和观点。《科学》，348(6239):1130–1132。

Delia Baldassarri 和 Andrew Gelman。2008年。无约束的党派人士：美国公众舆论中的政治极化与趋势。《美国社会学杂志》，114(2):408–446。

Ramy Baly、Giovanni Da San Martino、James Glass 和 Preslav Nakov。2020年。我们可以检测你的偏见：预测新闻文章的政治意识形态。*arXiv* 预印本 *arXiv:2010.05338*。

William D Berry、Evan J Ringquist、Richard C Fording 和 Russell L Hanson。2007年。州公民意识形态的测量与稳定性。《州政治与政策季刊》，7(2):111–132。

Alberto Bisin 和 Thierry Verdier。2000年。文化传递、投票与政治意识形态模型。《欧洲政治经济学杂志》，16(1):5–29。

AdamBonica。2014年。绘制意识形态市场图谱。《美国政治学杂志》，58(2):367–386。

Ian Budge、Hearl Derek、David Robertson、Derek Hearl 等。1987年。意识形态、战略与政党变迁：对*19*个民主国家战后选举纲领的空间分析。剑桥大学出版社。Tianqi Chen 和 Carlos Guestrin。2016年。XGBoost：一种可扩展的树提升系统。载于第*22*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第785–794页。Raviv Cohen 和 Derek Ruths。2013年。在Twitter上对政治倾向进行分类：这并不容易！载于*AAAI*网络与社交媒体国际会议论文集，第7卷。Michael D Conover、Bruno Gonçalves、Jacob Ratkiewicz、Alessandro Flammini 和 Filippo Menczer。2011年。预测Twitter用户的政治立场。载于*2011*年*IEEE*第三届隐私、安全、风险与信任国际会议及*2011*年*IEEE*第三届社会计算国际会议，第192–199页。IEEE。Maureen A Craig 和 Jennifer A Richeson。2014年。站在“多数—少数”美国的边缘：种族人口结构变化所引发的感知到的地位威胁会影响美国白人的政治意识形态。《心理科学》，25(6)：1189–1197。Hal Daumé III。2009年。令人沮丧地简单的领域自适应。*arXiv*预印本 *arXiv:0907.1815*。Michael X Delli Carpini。2000年。Gen. com：青年、公民参与与新的信息环境。《政治传播》，17(4)：341–349。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018年。BERT：面向语言理解的深度双向变换器预训练。*arXiv*预印本 *arXiv:1810.04805*。

Rahul Dey 和 FathiM Salem。2017 年。门控循环单元（GRU）神经网络的门控变体。载于*2017* 年 *IEEE* 第*60*届国际中西部电路与系统研讨会（*MWSCAS*），第1597–1600页。IEEE。

Nat Dilokthanakul、Pedro AM Mediano、Marta Garnelo、Matthew CH Lee、Hugh Salimbeni、Kai Arulkumaran 和 Murray Shanahan。2016 年。利用高斯混合变分自动编码器进行深度无监督聚类。*arXiv* 预印本 *arXiv:1611.02648*。

Guodong Gao、Brad N Greenwood、Ritu Agarwal 和 Jeffrey S McCullough。2015。发声的少数派与沉默的大多数。*MIS* 季刊，39（3）：565–590。

Matthew Gentzkow 和 Jesse M Shapiro。2010 年。是什么推动了媒体倾向？来自美国日报的证据。《计量经济学》，78（1）：35–71。

Sean M Gerrish 和 David M Blei。2011年。根据文本预测立法表决记录。载于第*28*届国际机器学习会议（*ICML 2011*）论文集。Xavier Glorot、Antoine Bordes 和 Yoshua Bengio。2011年。面向大规模情感分类的领域自适应：一种深度学习方法。载于*ICML*。Geoffrey Hinton、Nitish Srivastava 和 Kevin Swersky。2012年。机器学习神经网络讲座 6a：小批量梯度下降概述。引文，14(8):2。Michael H Hunt。2009年。意识形态与美国外交政策。耶鲁大学出版社。Mohit Iyyer、Peter Enns、Jordan Boyd-Graber 和 Philip Resnik。2014年。运用递归神经网络检测美国政治意识形态。载于计算语言学协会第*52*届年会论文集（第*1*卷：长篇论文），第1113–1122页。John E Jackson 和 John W Kingdon。1992年。意识形态、利益集团评分与立法投票。美国政治科学杂志，第805–823页。Diederik P Kingma、Shakir Mohamed、Danilo Ji menez Rezende 和 Max Welling。2014年。基于深度生成模型的半监督学习。载于神经信息处理系统进展，第3581–3589页。Diederik P Kingma 和 Max Welling。2013年。自动编码变分贝叶斯。*arXiv* 预印本 *arXiv:1312.6114*。Andrew Kohut 和 Carol Bowman。2018年。美国政治中的发声少数派。载于《广播——被遗忘的媒介》，第45–58页。劳特利奇。

Matthew S Levendusky 和 Jeremy C Pope。2010年。测量总体层面的意识形态异质性。立法研究季刊，35(2):259–282。

Steven D Levitt. 1996年。参议员如何投票？厘清选民偏好、党派归属和参议员意识形态的作用。《美国经济评论》，第425–441页。

Shuhua Liu 和 Patrick Jansson。2017。利用词嵌入和主题模型可视化，从 Instagram 数据中识别城市事件。工作论文。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019年。RoBERTa：一种经过稳健优化的 BERT 预训练方法。*arXiv* 预印本 *arXiv:1907.11692*。

Dylen Lysen、Laura Ziegler 和 Blaise Mesa。2022。堪萨斯州选民决定维持堕胎合法，否决一项修正案。*NPR* 新闻。

Colleen McClain. 2021. 70%的美国社交媒体用户从不或很少发布或分享有关政治和社会议题的内容。 皮尤研究中心。

Wendy W Moe、David A Schweidel 和 Michael Trusov。2011。哪些因素影响顾客的在线评论。《麻省理工学院斯隆管理评论》，53(1):14。

Negar Mokhberian、Andrés Abeliuk、Patrick Cum-mings 和 Kristina Lerman。2020年。新闻的道德框架与意识形态偏见。载于国际社会信息学会议，第206–219页。施普林格。EniMustafaraj、Samantha Finn、Carolyn Whitlock 和 Panagiotis T Metaxas。2011。发声的少数派与沉默的大多数：发现长尾群体的观点。载于*2011 IEEE* 第三届隐私、安全、风险与信任国际会议及 *2011 IEEE* 第三届社会计算国际会议，第103–110页。IEEE。David C Nice。1985年。州级政党意识形态与政策制定。《政策研究杂志》，13(4):780。Avital Oliver、Augustus Odena、Colin Raffel、Ekin D Cubuk 和 Ian J Goodfellow。2018。深度半监督学习算法的真实评估。*arXiv* 预印本 *arXiv:1804.09170*。Jonathan Ortigosa-Hernández、Juan Diego Rodríguez、Leandro Alzate、Manuel Lucania、Inaki Inza 和 Jose A Lozano。2012a 年。通过多维分类器的半监督学习开展情感分析。《神经计算》，92:98–115。Jonathan Ortigosa-Hernández、Juan Diego Rodríguez、Leandro Alzate、Manuel Lucania、Inaki Inza 和 Jose A Lozano。2012b年。通过多维分类器的半监督学习开展情感分析。《神经计算》，92:98–115。

Ferran Pla 和 Lluís-F Hurtado。2014。使用情感分析技术识别 Twitter 上的政治倾向。载于 *COLING 2014*论文集：第*25*届国际计算语言学会议：技术论文，第183–192页。

Keith T Poole 和 Howard Rosenthal。2001。D-nominate 十年后：对《国会：唱名表决的政治经济史》的比较性更新。《立法研究季刊》，第5–29页。

Niklas Potrafke。2018。美国的政府意识形态与经济政策制定——一项调查。《公共选择》，174(1):145–207。

Daniel Preo,tiuc-Pietro、Ye Liu、Daniel Hopkins 和 Lyle Ungar。2017。超越二元标签：Twitter用户的政治意识形态预测。载于第*55*届计算语言学协会年会论文集（第*1*卷：长篇论文），第729–740页。

Juan José Rodriguez、Ludmila I Kuncheva 和 Car- los J Alonso。2006。旋转森林：一种新型分类器集成方法。《*IEEE*模式分析与机器智能汇刊》，28(10):1619– 1630。

Thomas J Rudolph。2009。政治信任、意识形态与公众对减税的支持。《*Public Opinion Quarterly*》，73(1):144–158。

Suzanna Sia、Ayush Dalmia 和 Sabrina J Mielke。2020年。厌倦主题模型了吗？预训练词嵌入的聚类也能快速生成优质主题！*arXiv* 预印本 *arXiv:2004.14914*。Jafar Tanha 、 Maarten van Someren 和 Hamideh Af- sarmanesh。2017。用于决策树分类器的半监督自训练。《国际机器学习与控制论杂志》，8(1):355–370。Jakub Tomczak 和 Max Welling。2018。采用 VampPrior 的 VAE。载于人工智能与统计国际会议，第1214–1223页。PMLR。Stefan Wojcik 和 Adam Hughes。2019年。Twitter 用户规模分析。皮尤研究中心，24。Zhiping Xiao、Weiping Song、Haoyan Xu、Zhicheng Ren 和 Yizhou Sun。2020年。Timme：通过多任务多关系嵌入进行 Twitter 意识形态检测。载于第*26*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第2258–2268页。Hao Yan、Sanmay Das、Allen Lavoie、Sirui Li 和 Betsy Sinclair。2019年。国会分类挑战：领域特异性与党派倾向强度。载于*2019*年*ACM*经济学与计算会议论文集，第71–89页。Hao Yan、Allen Lavoie 和 Sanmay Das。2017。根据文本分类政治倾向的风险。载于 *LINKDEM@ IJCAI*。

David Yarowsky。1995。无监督词义消歧可与有监督方法相媲美。载于第*33*届计算语言学协会年会，第189–196页。

Bei Yu、Stefan Kaufmann 和 Daniel Diermeier。2008。根据政治演讲对政党归属进行分类。《信息技术与政治》，5(1):33–48。

Pengyuan Zhang、Yulan Liu 和 Thomas Hain。2014。会议识别中的半监督深度神经网络训练。载于*2014 IEEE* 口语技术研讨会（*SLT*），第141–146页。IEEE。

Joshua N Zingher。2014。美国各州针对无证移民的法律之意识形态与选举决定因素。《州政治与政策季刊》，14(1):90–117。

Yang Zou、Zhiding Yu、BVK Kumar 和 Jinsong Wang。2018。通过类别均衡的自训练实现语义分割的无监督域适应。载于《欧洲计算机视觉会议（*ECCV*）论文集》，第 289–305 页。

A 附录

A.1 数据集详情。

国会演讲 是一个语料库，收录了国会众议院或参议院议员在 2009-2020 年间发表的（22 万篇）演讲的文字记录，涵盖奥巴马和特朗普两届总统任期。每篇演讲均提供演讲者 ID、年份、演讲者的政党归属以及演讲标题。每篇演讲文档的意识形态标签由其演讲者的政党归属确定。尽管这种代理指标似乎并不完美，但已有研究表明，政党归属与国会议员的意识形态总体上高度一致（Levitt，1996年）。文档/发言者的极端程度可根据 DW-NOMINATE 分数确定。

枪支辩论论坛是一个在线论坛（ debatepolitics.com）上的语料库，包含 6 万条围绕枪支暴力和枪支管制议题的帖子。在该论坛发帖的每位用户都可以选择归属以下意识形态群体之一：略偏自由派、略偏保守派、自由派、保守派、极端自由派、极端保守派、温和派、中间派、自由意志主义者、无政府主义者和进步派。本研究仅关注自由派—保守派光谱，包括自我认同为略偏自由派、略偏保守派、自由派、保守派、极端自由派、极端保守派和温和派的用户所发的帖子（图。4）。这样一来，所有帖子及其作者都被置于七点量表上（Preo₀tiuc- Pietro 等人，2017）。

DW-NOMINATE 是一个数据集，用于描述众议院和参议院议员根据其投票记录得出的意识形态得分，数据来自 (voteview.com)。得分（主要维度）在自由派（-1）到保守派（+1）之间连续变化，并能解释大多数（>80%）的投票差异。有关如何根据投票数据计算得分的详细信息，请参阅 (Pooleand Rosenthal，2001)。DW-NOMINATE 得分被广泛视为基准指标 (Jackson 和 Kingdon，1992)。

A.2 基线方法与实验详情

8l-DNN 一个具有 7 个中间层和一个大小为 1 的输出层的全连接网络。对于 GloVe 嵌入，输入文档嵌入通过对维度为 300 的词嵌入取平均值获得；对于 BERT，则使用维度为 1024 的 CLS（或池化）嵌入来表示输入文档。中间层的大小依次为 800、800、800、400、250、800 和 800

（为简洁起见，此类全连接网络的结构将记为(800, 800, 800, 400, 250, 800, 800, 1)，下文其他全连接网络也采用相同记法）。除输出层使用Sigmoid 函数外，我们均使用 ReLU 作为激活函数。训练时的学习率设为 0.001，在最后两层使用 0.01 的 L2 正则化，并使用 RMSProp 进行优化（Hinton 等人，2012）。此后，8l-DNN 将作为其他一些基线模型的构建模块。除非另有说明，均采用上述相同的结构和参数。

GS 模型源自一种基准方法，该方法用于衡量文本中的政治倾向，由 Gentzkow 和 Shapiro 于 2020年开发（Gentzkow 和 Shapiro，2010年）。该方法基于朴素贝叶斯假设。我们按照（Gentzkow 和 Shapiro，2010年）中的描述，选出极化程度最高的短语，然后对词频进行政党标签回归，再将这些极化短语的系数按短语频率加权求和，从而得到测试数据中每篇演讲的政党倾向。

门控循环单元 该模型将文本作为输入嵌入序列（Dey 和 Salem，2017），并输出一个长度为 300 的向量。随后将该输出向量输入 Dropout 层（p =0.5），再输入一个结构为（800、800、400、1）的全连接网络。除最后一层使用 Sigmoid 激活函数外，所有层均使用 ReLU 激活函数。训练该门控循环单元时，我们使用 Adam 优化器，初始学习率为 0.01

BERT/RoBERTa 模型BERT 和 RoBERTa 模型都会为句子输出两种表示——池化/CLS 嵌入或嵌入序列（Devlin 等人，2018；Baly 等人，2020年；Liu 等人，2019年）。我们在实验中尝试了这两种表示方式，但只报告了 RoBERTa 的序列表示结果，因为其性能最佳。嵌入序列作为输入送入一个网络，其结构与门控循环单元部分所述的结构相同。对于 RoBERTa-SBBG 模型，嵌入序列首先作为输入送入单个门控循环单元层，以生成维度为 300 的嵌入；随后，这些嵌入被视为输入样本，送入 SBBG 模型部分所述的框架[7]。与 GloVe 不同，我们允许

[7]由于 BERT/RoBERTa 嵌入本质上是非线性的，由所有上下文信息的加权和构成，因此与作为 BBBG 核心的文档线性分解模型相冲突，我们只对 RoBERTa 进行了 SBBG 消融实验，而没有使用 BBBG 完整模型。不过，正如我们的实验所示，与 GloVe 相比，改用 RoBERTa 对提升学习性能几乎没有帮助。

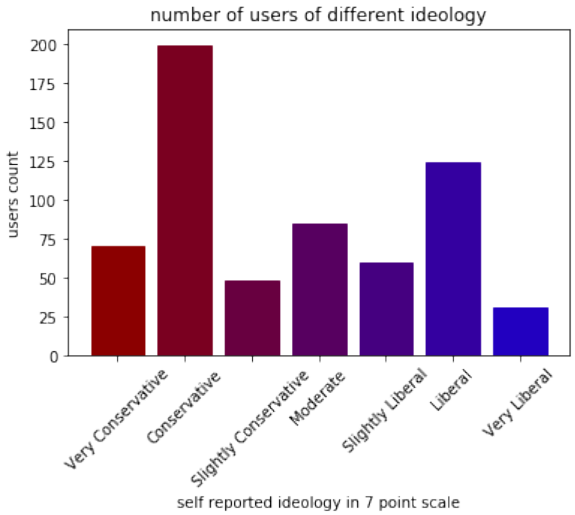


图4：各意识形态群体的发帖者数量。可以看到，保守派或极端保守派发帖者约占总数的 45%，而自由派或非常自由派发帖者占 25%。在轻度倾向的发帖者中，情况则恰好相反。他们中略偏自由派的多于略偏保守派的。

在训练期间对嵌入进行微调。

领域自适应领域自适应模块基于 RoBERTa 生成的序列嵌入。在训练阶段，带标签文本被分配至源领域，数量相同的随机抽取的无标签文本被分配至目标领域。在推理阶段，报告标签分类器在测试数据上的预测结果。

标签传播 和**自训练** 都是典型的半监督学习（SSL）模型，用于处理训练数据标注不足的情况（Yarowsky，1995；Ortigosa-Hernández 等人，2012b年）。尽管监督信息不足是意识形态学习中的主要障碍，但半监督学习在这一领域很少被采用；本研究将其纳入实验，以探究它为何未被广泛用于解决这一问题。在我们的实验中，标签传播采用 kNN（k=7）核函数，自训练则采用 8l-DNN、随机森林或 XGBoost 作为核函数（Rodriguez 等人，2006；Chen 和 Guestrin，2016）。

A.3 BBBG训练

根据公式3，BBBG模型的生成部分可视化如图5所示，而整个模型如图6所示。它包含四个主要组成部分，具体如下。

x到z。VAE框架的编码器通过两个全连接（FC）网络，从$\mathbf{x}$ 中学习$\mathbf{z}$(dim=50)的均值和标准差。$\mathbf{z}$ 通过

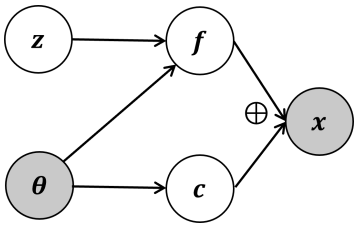


图5：BBBG 生成部分的示意图。

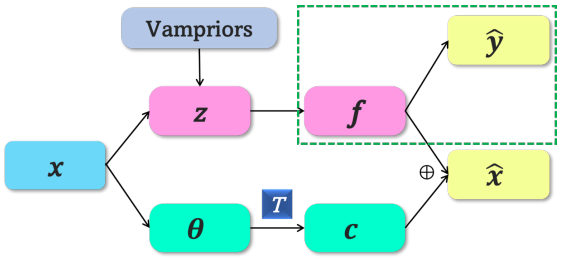


图6：采用 VampPriors 的双分支 BBBG 网络。此处每个箭头表示多个层（>3），不包括重构层。绿色方块突出显示了意识形态监督模块。

重参数化技巧进行采样。用于学习**z** 均值的FC网络形状为 (800, 800, 800, 400, 50)。用于学习方差的FC网络形状相同，并且与均值网络在前两个中间层共享权重。除方差FC网络的最后一层外，所有层均使用 ReLU 激活函数；该层使用阶跃函数，将所有介于 -6 到 3 之间的输入映射为 1，其他输入映射为 0。

z到y。意识形态预测单元由一个形状为 (800, 800, 1) 的全连接网络构成。该网络使用 ReLU 作为激活函数，最后一层则使用 Sigmoid。

$\tilde{}$x 到 θ 到c。这是中性上下文分支。$\tilde{}$主题分配向量 θ 通过一个形状为 (800, 1600, 400, 68) 的全连接网络从 **x** 中学习得到，其中 68 是主题总数（包括“其他”）。上下文向量**c** 通过$\tilde{}$将主题矩阵 T 与 θ 相乘得到（见下文说明）。

$\tilde{}$z 和 θ 到f。变分自编码器框架的解码器用于解码**f**。首先将**z** 和 θ 拼接。随后将所得向量作为形状为 (800, 800, 300) 的全连接网络的输入$\tilde{}$，以解码**f**。接下来，通过将**f**和**c**相加来重构**X**。

从伪输入生成 **z** 的先验分布（VampPriors），遵循（Tomczak 和 Welling，2018）中的方法。

超参数 对于国会报告上的实验，我们在总样本的 10% 上调优主模型的超参数（单分支版本沿用相同的一组超参数），并将其用作验证集。我们使用了 0.001 的学习率和 RMSProp 优化器。其他参数包括 VampPriors 的采样所依据的高斯分布的均值（0.02）和标准差（0.3）（括号中的数值表示最终调优值）（Tomczak 和 Welling，2018）、**z** 先验分布的 KL 散度退火参数（15）（Dilokthanakul 等人，2016），以及训练轮数（实验1为 10 轮，实验2为 15 轮）。意识形态预测分支采用了 0.01 的 L2 正则化。在 Debatepolitics 语料库上的实验沿用了上述相同的超参数。

$\tilde{}$SBBG 与 BBBG 相同，但不包含从 **x** 到 θ到 **c**，此外，**x** 直接由**z**重建，使用形状为 (800, 800, 300) 的全连接网络。

A.4 主题矩阵初始化，T 和主题分配 θ

虽然可以通过潜在狄利克雷分配（LDA）等无监督模型获得主题分配，但我们可能会遇到两个问题。首先，LDA模型可能会发现非中性的关联。其次，纯粹基于词语分布的模型可能无法发现不常见的主题。在国会报告中，主题主要集中在财政支出、医疗保健、经济和政治上（>60%）。一些重要但不常见的主题，例如枪支（<1%）和堕胎（<2%），很容易被 LDA模型忽略。为了解决这些问题，我们采用了一种流行的嵌入相似度方法（Liu 和 Jansson，2017；Sia 等人，2020年）。首先，我们通过研读大量国会报告（5K）和 5k 份法案摘要，手动选取了 67 个主要主题；对于每个主题，我们构建了一组中性词，请专家提出 5 到 10 个中性种子词（例如，枪支主题的“枪械”和“枪支”），再根据词嵌入空间中的邻近关系将词集扩充至 100 个词；接着，我们手动检查每个种子词及扩充词，剔除不相关、罕见或“非中性”的词（即明显带有党派色彩的词，例如医疗保健主题中的“奥巴马医改”、LGBTQ 主题中的“有罪的”，以及堕胎主题中的“杀婴”）；随后，将所有选定词的嵌入取平均值，计算出主题的嵌入；最后，将上述所有主题向量按顺序堆叠，形成初始主题矩阵 T_0，矩阵的每一行对应某个主题的主题向量，之后将

其用于初始化主题向量矩阵 T（下一节将介绍另一种通过统计推断确定中性词集的方法）。如果我们知道文档 x的主题分配，那么上下文向量为$c = T\theta$。对于 Debatepolitics，主题分配是已知的。对于国会报告，主题分配通过计算文档嵌入与上述上下文向量之间的余弦相似度来初始化（Liu 和 Jansson，2017；Sia 等人，2020年），并稍作修改。我们没有进行硬分配，而是对一个文档的嵌入与各上下文向量的相似度应用 Softmax 转换，生成初始软分配 θ_0 。如果最大余弦相似度低于某个阈值（约为所有最大值的第 10 个百分位数），我们就将主题标记为“其他”。

随后，我们在训练过程中使用这些初始主题分配（如正文所述），从而使 θ 和 T 能够通过 x进行更新和微调。如图 6 所示，θ 通过前馈网络从 x 中学习得到，而 c 则通过 $c = T\theta$获得。此处，T 以 T_0的值初始化，且 θ 和 T 均为可训练变量。在损失函数中，我们加入了两个额外项，以最小化 θ （或 T）与 θ_0 （或 T_0）初始值之间的均方差。这限制了对 θ和 T局部最优解的搜索范围，使其位于各自原始值的邻域内。

训练完成后，我们对更新后的中性语境向量矩阵 T进行了事后验证：1）检查其是否与各主题对应的极化轴正交（见正文）；2）人工检查原始嵌入空间中各主题上下文向量的邻域。对于每个主题，我们收集了嵌入空间中距离上下文向量最近的 20 个词。

A.5 获取中性上下文向量中中性词的另一种方法

本节将给出主题层面中性词的统计定义，这种定义可作为初始化中性上下文向量的另一种方式。考虑选择一个主题，例如“枪支”，令 $\mathcal{W}_{gun}^{D}$ 为民主党人通常用于讨论枪支的词语或短语集合，令 $\mathcal{W}_{gun}^{R}$ 为词语或短语集合共和党人通常使用的**词语** $\in \mathcal{W}_{gun} \triangleq \mathcal{W}_{gun}^{D} \cup \mathcal{W}_{gun}^{R}$，我们计算了其文档频率 fr_{doc}，以及不同党派比例之间文档频率的差异 $\Delta fr \triangleq |fr_{doc}^{D} - fr_{doc}^{R}|$（或者，也可以使用$\chi^2$ 每个词语的Gentzkow 和 Shapiro，2010年）。满足以下条件的词语 $\mathbf{w}$，其 $fr_{doc}(\mathbf{w}) \leq \alpha$ 和 $\Delta fr(\mathbf{w}) \leq \beta$将从 $\mathcal{W}_{gun}^{D}$ 和 $\mathcal{W}_{gun}^{R}$中剔除。最后，枪支主题的中性词集合定义为$\mathcal{W}_{gun}^{\mathcal{N}} \triangleq \mathcal{W}_{gun}^{D} \cap \mathcal{W}_{gun}^{R}$。该中性集合包含“枪支”“枪支”“弹药”等词语。

根据这一定义，明显受意识形态驱动的词语（如“libtards”“杀婴”“death (panel)”等）会被剔除。此外，中性词集合还排除了与特定主题相关、且受到某一意识形态群体偏好的词语，例如移民主题中的“illegal”和“undocumented”。这样一来，残差就能尽可能多地捕捉信息，从而正确推断意识形态。这一方法受到Gentzkow 和 Shapiro的论文启发。他们采用了类似的流程，但与我们不同的是，他们选择了χ^2 值最高的意识形态词语。

A.6 关于屏蔽的补充细节

屏蔽过程通过随机抽样，或选取极端程度最高的 X%（X 从 80 到 1）的民主党人和共和党人并将其设为未屏蔽来完成。这里的极端程度根据国会议员的 DW-NOMINATE 分数确定，该分数被视为衡量政治意识形态的基准指标（Jackson 和 Kingdon，1992）。我们将前一种监督称为无偏监督（因为样本中会同时包含极端和非极端成员），后一种则称为有偏监督。

对于第二个实验，文本表示为词嵌入的方式和屏蔽流程与实验1相同。不过，鉴于意识形态真实值并非连续表示，而是采用七点量表，且各组用户和帖子的分布并不均衡，我们采用了**不放回**加权抽样方案，以模拟观察到的较不极端群体中的结果稀缺性，具体如下。

在加权抽样方案中，与常规群体相比，极端群体（即极端自由派或保守派）发布的帖子被抽入未屏蔽部分的可能性高出 10 倍；

而常规群体（即自由派或保守派）发布的帖子被抽入未屏蔽部分的可能性，又比轻微倾向群体（略偏自由派或略偏保守派）高出 20 倍。这样一来，当未屏蔽样本相较于总人口较为稀少（不足 8%）时，它们将主要由极端发帖者发布的帖子构成。即使未屏蔽样本的覆盖率达到总人口的 80%，其中仍极不可能包含来自轻微倾向用户的样本。这一方案模拟了现实世界中的情形：立场越极端的个体越积极地表达政治观点，也越愿意透露自己的意识形态；而“沉默的大多数”主要是温和派或略有倾向者，在政治上相对不爱发声，其政治观点和意识形态标签大多未被观察到。

A.7 国会报告的全体与党派主题

所有主题的名称包括“IP”、“信息技术”、“堕胎”、“学术”、“农业”、“商业”、“儿童”、“中国”、“商业”、“犯罪”、“文化”、“国土安全”、“被拘留者”、“弱势群体”、“灾害”、“种族与少数群体”、“疾病”、“成瘾性药物”、“经济”、“教育”、“环境”、“家庭”、“联邦运作”、“金融”、“财政主题”、“食品”、“枪支”、“健康主题”、“医疗保健”、“高科技”、“住房”、“移民”、“工业”、“网络安全”、“基础设施”、“国际事务”、“伊朗、叙利亚、利比亚”、“伊拉克”、“以色列”、“陪审团”、“LGBTQ”、“媒体”、“军事综合体”、“原住民”、“核能”、“警察”、“政治行动”、“邮政”、“研发”、“宗教”、“可再生能源”、“储备”、“俄罗斯”、“安全”、“体育”、“税收”、“恐怖主义”、“贸易”、“传统能源”、“交通运输”、“退伍军人”、“越南”、“投票”、“废弃物”、“福利”、“女性”、“劳动力”和“其他”。

用于验证上下文向量中立性的两极化主题（不包括出现次数少于 500 次的罕见主题）包括“堕胎”“农业”“被拘留者”“弱势群体”“灾害”“种族与少数群体”“疾病”“成瘾性药物”“经济”“环境”“财政”“枪支”“健康主题”“医疗保健”“移民”“国际”“伊朗、叙利亚、利比亚”“伊拉克”“以色列”“军事综合体”“可再生能源”“俄罗斯”“传统能源”“福利”“劳动力”。

A.8 众包实验的更多详情

为验证从 BBBG 学到的上下文向量在多大程度上符合人类对中立性的判断，我们通过 Prolific 平台招募了 476 名参与者，他们均为美国公民。其中，大多数人居住在加利福尼亚州（67 人）、纽约州（40 人）和得克萨斯州（34 人）。受访者中，50% 为女性，47% 为男性。约 55% 的参与者拥有学士或更高学位，34% 持有高中毕业文凭，10% 拥有社区学院学位。至于职业，超过四分之一的参与者表示目前失业。在有工作的参与者中，大多数从事管理及专业类工作（129 人），其次是服务业（89 人）和销售业（53 人）。根据参与者在七点评分量表上的自我意识形态报告，大多数（62%）支持民主党观点，而支持共和党观点的参与者和温和派／中间派各占 19%。

我们提取了六个热门党派主题（堕胎、枪支、医疗保健、移民、社会福利和女性权利）的上下文向量，并计算了这些上下文向量与词库中排名前 10K 的每个词的投影嵌入之间的余弦距离（1 - 余弦相似度）。每个主题中距离最近的 10 个词里，有 5 个被选作语境邻近 词。作为对照，我们从知名党派新闻媒体中手动搜集了五个具有典型保守派／自由派特征的词语／短语（稍后我们会介绍每个词语／短语的来源）。这些词语／短语将称为参照词。因此，每个主题将调查 15 个词语／短语（共 75 个）。

调查期间，每位受访者都需要回答若干人口统计问题，包括其自身的意识形态倾向（自由派、保守派、中间派／两者皆非），并为随机抽取的 30 个目标域或参照词语／短语评分。具体而言，我们请他们评估：1）他们认为这些词语与各个主题的相关程度（李克特量表 1 至 7 分）；2）他们认为这些词语的意识形态倾向偏向自由派还是保守派（1 分表示非常自由派，7 分表示非常保守派；此外还提供了“不知道／不确定”选项，以提高评分准确性）。为避免主观偏差（例如，自由派受访者倾向于给自由派词语打更接近中立的分数），我们对自由派受访者针对每个词语／短语给出的结果进行了随机下采样（

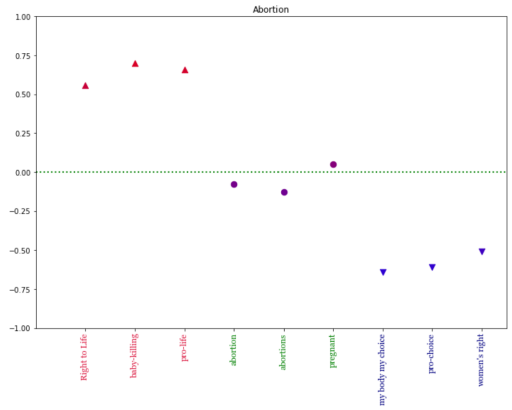


图7：众包工作者评定的堕胎主题平均意识形态倾向得分。目标词语/短语、自由派词语/短语和保守派词语/短语分别以绿色、蓝色和红色标示。纵轴为意识形态倾向得分。每个点的颜色也对应其得分值。

因为 Prolific 上的自由派众包工作者人数约为保守派或中间派的两倍）。最终，每个词语／短语都由数量大致相同的自由派、保守派和中间派受访者评分（平均约 65 人）。分析期间，我们将相关性得分重新缩放至 [0,1]，将意识形态倾向得分重新缩放至 [-1, 1]。

我们计算了每个词语的意识形态倾向得分，并将其汇总至主题层面（见表8。图7展示了堕胎主题中受调查词语/短语的意识形态倾向得分（为清晰起见，5 个词语/短语中选取了 3 个进行展示）。

A.9 意义建构：它能捕捉意识形态吗?

如果**z** 本质上具有意识形态属性，那么就应存在一些主轴，沿着这些轴，常见意识形态群体（政党标签）中的成员会彼此区分。图3a 展示了所有作者的意识形态**z** 的前两个主成分（按作者取均值后汇总），颜色对应作者的政党归属（共和党人为红色；民主党人为蓝色）。除少数例外，成员沿第二主成分明显分离，这表明**z** 确实是一种意识形态表征。

如果我们的模型能够准确表征并估计作者的真实意识形态，那么按作者汇总的输出均值应与通过独立于国会演讲的数据获得的传统意识形态分数高度相关，例如基于投票行为的*DW-NOMINATE*分数。图3 展示了我们的模型预测的倾向得分（即从**z**预测意识形态标签的机制所产生的输出）与作者的主要 DW-NOMINATE 分数之间的关系。两个分数高度相关（$R^2 \sim 0.9$），这表明我们的模型仅需最少的监督，就准确估计了文档作者的意识形态。在确认模型有效后，我们现在转向评估标签稀缺和极端偏差情况下的表现。

B 补充结果

B.1 实验1的补充表格

表6 展示了在国会报告语料库上进行的实验中，BBBG 与主论文中提及的其他几种基线方法之间更多的对比结果。这旨在证明，主论文中汇总的方法优于此处报告的方法。

B.2 消融实验结果

我们对主模型进行了一些消融实验。首先是单分支模型（SBBG），该模型移除了语境学习分支。该模型在两个基准数据集上的结果见表1和表2。如表1 和表2所示，对于所有数据集，无论监督信号存在偏差还是不存在偏差，BBBG 的表现均优于 SSSG。此外，稀缺性越高，二者的表现差异越大。由于 BBBG 与 SSSG 的性能差异可归因于文档分解，这表明将文档分解为中性语境和位置向量能够提高准确率，并增强模型对领域偏移的泛化能力。

我们还测试了 K 取不同于 2 的值（如 1 或 3）时的影响。请注意，当 K=1时，模型的变分部分退化为采用单峰高斯先验分布的普通变分自编码器。表7展示了结合不同 K 值（1 或 3）与 SBBG 或 BBBG 的模型实验结果（例如，SBBG_K1 是 K=1）。为便于比较，我们还纳入了主模型 BBBG 及其单分支变体 SBBG，这两者的 K 值均为 2。如表7所示，当 K 值偏离 2 时，模型性能会下降；与 SBBG 结合则会进一步降低性能。

表5：Prolific 调查实验中选取的参考词和目标词语／短语。

	参考 保守派_	参考 自由派_	语境_邻域
堕胎	生命权、杀婴、不道德， 部分分娩堕胎，支持生命	堕胎权，支持选择， 我的身体，我做主， 生育权，女性权利	堕胎，堕胎，分娩， 妊娠，怀孕
Gun	美国宪法第二修正案，持枪权， 守法公民，自卫 拥有枪支的权利	枪支禁令，枪支安全，大规模枪击事件， 校园屠杀，严格的枪支管制	枪械，枪支，枪支， 枪击、武器
医疗保健	死亡委员会，强制登记 掌握在.....手中的医疗保健 政府官僚， 更高的保费，增税	可负担的医疗服务，全民医疗保健 更容易获得预防性医疗服务和 初级卫生保健，扩大医疗补助计划 全民医疗保健	健康维护组织， 医疗保险覆盖、医疗补助计划， 保费、未参保者
移民	外国籍罪犯、修建隔离墙， 贩运者、非法移民 跨境入侵者	儿童笼舍、移民合法化， 暂停驱逐出境， 入籍途径、无证移民	移民归化局（INS）， 外籍人士、移民、难民、签证
福利	懒人的福利、社会负担， 机会社会， 臃肿政府、福利女王	冠状病毒救济方案， 面向所有劳动者的带薪家庭假／病假， 社会安全网、普惠儿童照护， 全民基本收入（UBI）	收入、收入、保费， 薪水、工资
女性权利	家庭价值观、忠于婚姻， 服从丈夫，传统观念， 女性美德	针对女性的战争，同工同酬， 女权主义，女孩，我也是运动	婴儿，婴幼儿，妊娠， 青少年，女性

表 6：无偏或有偏监督下国会演讲数据的政党预测准确率，展示了其他基线方法与主模型之间的比较结果。最佳结果以**粗体** 显示，次佳结果带下划线。在标签稀缺时，BBBG 的表现显著优于大多数其他模型，并以蓝色标示。所示百分比为三次独立试验的平均值。

	无偏监督								有偏监督							
	80%	60%	40%	20%	8%	5%	3%	1%	80%	60%	40%	20%	8%	5%	3%	1%
ST-RF	70.0%	71.5%	71.1%	69.7%	69.7%	69.8%	72.8%	61.1%	74.3%	72.9%	68.9%	56.9%	61.3%	61.3%	61.3%	61.3%
BERT	67.4%	67.3%	66.4%	60.4%	53.8%	50.4%	50.0%	51.2%	64.7%	65.4%	66.1%	57.7%	64.8%	61.0%	61.0%	61.2%
RF	65.1%	64.8%	64.6%	64.6%	64.4%	63.6%	63.0%	61.2%	61.7%	57.5%	54.2%	53.0%	63.4%	61.7%	62.5%	61.3%
GRU	77.7%	77.3%	75.7%	72.4%	69.2%	67.5%	65.7%	60.3%	61.6%	57.3%	54.1%	52.9%	86.1%	73.0%	70.5%	66.6%
BBBG	**92.7%**	**92.9%**	**94.0%**	**93.2%**	**91.6%**	**89.8%**	**87.2%**	**81.2%**	**81.3%**	**81.3%**	**85.2%**	**83.3%**	**85.3%**	**77.3%**	**74.4%**	**71.5%**

表7：在有偏监督下，枪支辩论论坛数据的意识形态标签预测准确率，展示了主模型 BBBG 与其变体之间的比较结果。最佳结果以**粗体**显示，次佳结果以下划线标出。在标签稀缺时，BBBG 的表现显著优于大多数变体，相关结果以蓝色标示。所示百分比为三次独立试验的平均值。

	80%	60%	40%	20%	8%	5%	3%	1%
SBBG_K1	92.4%	90.2%	86.5%	81.5%	73.8%	69.9%	64.7%	60.4%
SBBG_K3	91.0%	89.0%	86.9%	82.3%	73.7%	67.7%	67.9%	59.5%
BBBG_K1	91.9%	90.0%	86.8%	83.8%	77.2%	70.4%	64.1%	58.4%
BBBG_K3	90.5%	88.0%	86.8%	83.3%	76.6%	64.4%	66.1%	59.2%
SBBG	92.2%	88.7%	88.4%	**86.1%**	73.0%	70.5%	66.6%	60.9%
BBBG	**94.3%**	**91.0%**	**90.2%**	85.3%	**77.3%**	**74.4%**	**71.5%**	**61.2%**

B.3 即使训练数据中缺少相关信息，BBBG 也能区分轻微倾向群体。

BBBG 和先前研究在监督充分的情况下都取得了显著的性能。因此，这项实验的一个关键发现是，在异质群体设置中，由于监督偏差，轻微倾向群体的数据更少；此时，表现良好的模型必须尤其擅长提取意识形态知识。

我们使用一种新颖的排名偏差（RD）指标比较了不同方案（见补充材料）。该指标将作者在意识形态光谱中的排名与中位排名进行比较。显然，轻微倾向子组的平均或中位 RD 小于极端子组的 RD（见图 8）。与 ST-DNN 相比，BBBG 能更好地区分极端子组和轻微倾向子组，且方差更低。

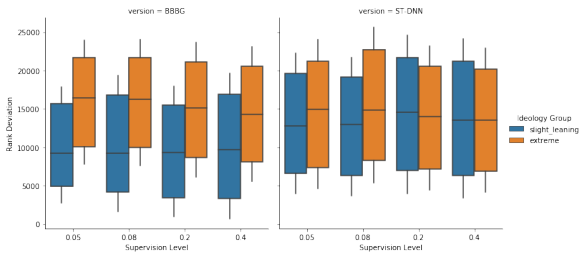


图 8：比较不同监督水平（X 轴上的百分比）下训练的 BBBG（左派）和 ST-DNN（排名）所对应的各个子组排名偏差分布。

B.4 众包实验的补充结果

按各主题对众包评定的意识形态倾向得分进行的详细比较见表。8。

表8：根据众包实验对词语的意识形态倾向进行评定，并按主题细分。我们请每位零工劳动者对随机抽取的词语进行评分，这些词语属于我们的目标邻域词或某个参照组（自由派词语或保守派词语）。得分重新缩放至 -1 到 1 的范围，其中 -1/1 分别对应极端自由派/极端保守派倾向。

	参照词		目标词
主题	保守派	自由派	邻域词语
堕胎	0.524*** (0.082)	-0.507*** (0.062)	0.039 (0.052)
经济	0.166 (0.074)	-0.341*** (0.073)	0.093 (0.05)
金融	0.047 (0.073)	-0.125*** (0.077)	0.058 (0.046)
Gun	0.498*** (0.067)	-0.338*** (0.079)	0.157 (0.067)
医疗保健	0.097* (0.081)	-0.413*** (0.06)	-0.0 (0.052)
移民	0.514*** (0.077)	-0.223** (0.083)	-0.041 (0.069)
Iraq	0.317*** (0.063)	-0.126*** (0.071)	0.108 (0.055)
可再生能源	0.246*** (0.074)	-0.439*** (0.053)	-0.038 (0.05)
福利	0.377*** (0.094)	-0.493*** (0.055)	0.124 (0.047)
女性	0.418*** (0.07)	-0.364*** (0.074)	0.08 (0.049)

我们目标词与任一参照组（自由派或保守派）之间群体评定的意识形态倾向差异的显著性通过双样本 t 检验计算。显著性水平标注在参照组后。* p<0.05；** p<0.01；*** p<0.001